PRED

Soutenance finale

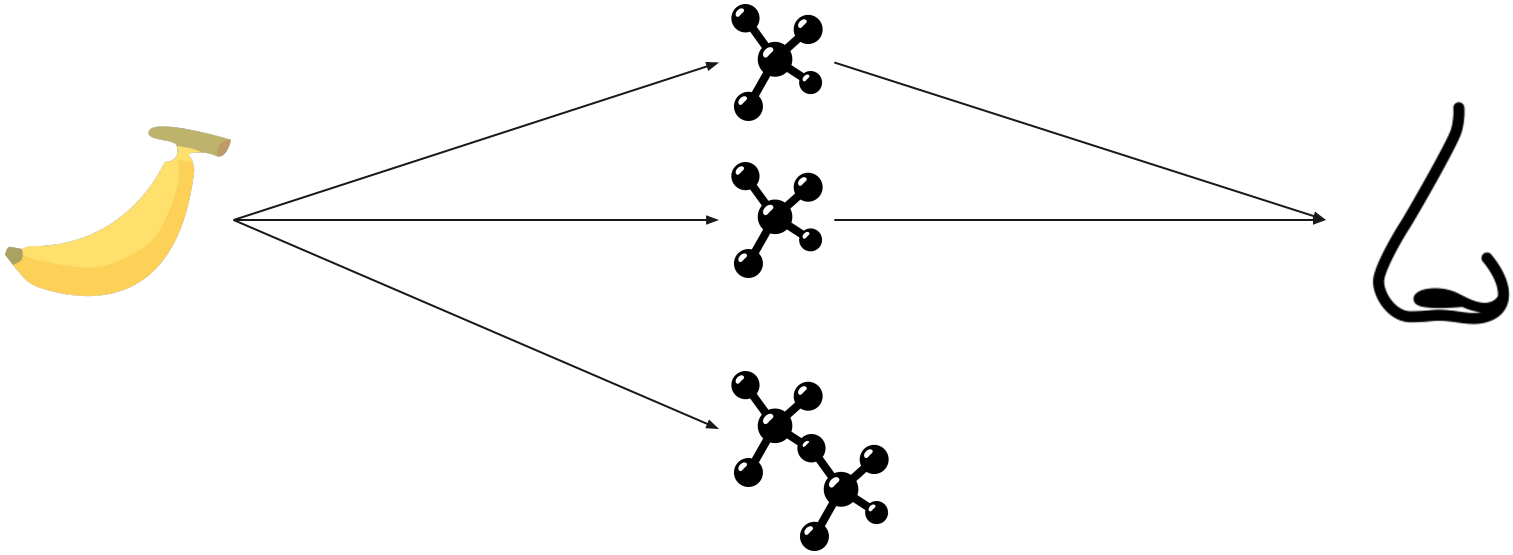
Fouille de données olfactives : clustering de molécules odorantes par Graph Neural Networks

Jolly Gabriel - Clouet Thomas

Sommaire

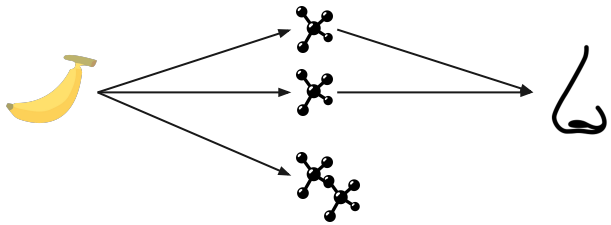
01. Présentation du sujet

Olfactométrie



Décomposition Tri Identification

Olfactométrie



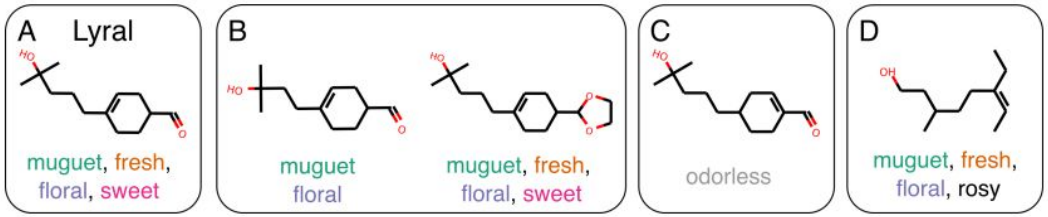


Figure 1 : Molécule et leurs descripteurs d'odeurs associés de l'article[1]

Problèmes
- On sent la molécule seule, ce qui n'est pas commun
- Nous percevons tous les odeurs de différentes façons

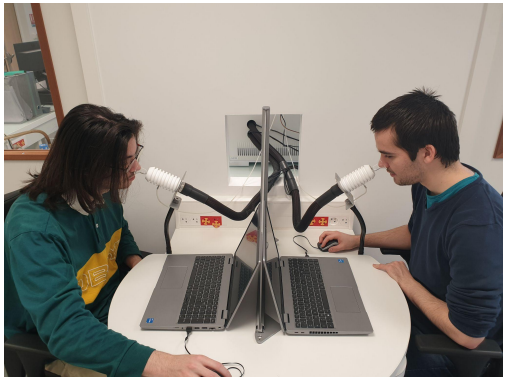
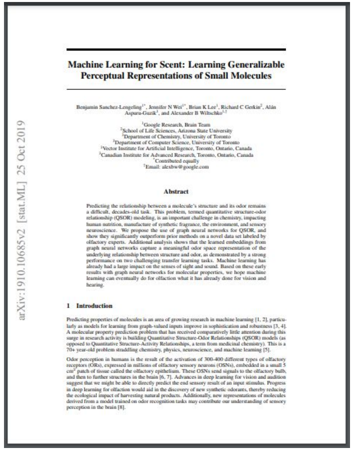


Figure 2 : Découverte de l'olfactométrie, au laboratoire d'Oniris

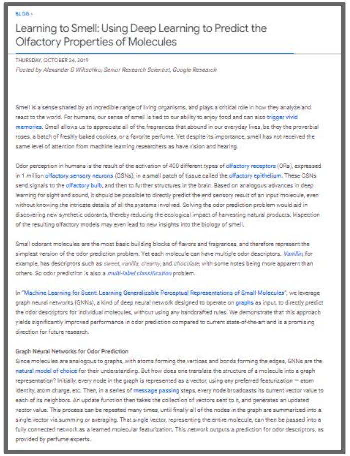
[1] https://arxiv.org/pdf/1910.10685.pdf

Nos objectifs : 1. Reproduire la démarche de l’équipe de Google brain, afin d'être en capacité de prédire les descripteurs d’odeur d’une molécule



Machine Learning for Scent: Learning Generalizable Perceptual Representations of Small Molecules

arXiv:1910.10685v2 [stat.ML] 25 Oct 2019



Learning to Smell: Using Deep Learning to Predict the Olfactory Properties of Molecules

THURSDAY, OCTOBER 24, 2019

Posted by Alexander B Wiltschko, Senior Research Scientist, Google Research

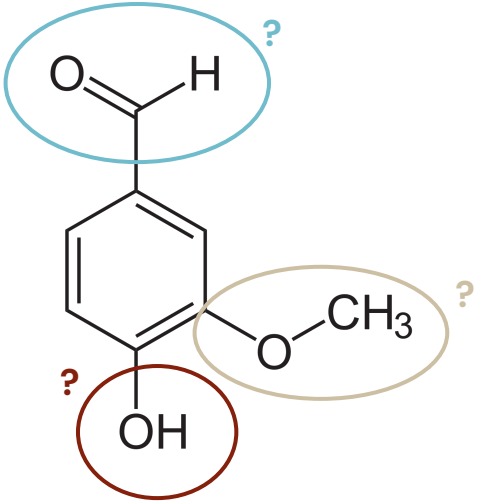
Équipe Brain de Google Research, 2019

Comment et pourquoi les GNNs sont utilisés pour la prédiction des odeurs

https://ai.googleblog.com/2019/10/learning-to-smell-using-deep-learning.html

https://arxiv.org/pdf/1910.10685.pdf

Nos objectifs : 2. Déterminer la partie de la molécule qui est émettrice du descripteur d'odeur



Molécule de vanilline

Lait frais

Chocolat

Crème

...

Descripteurs prédis

Nos plannings

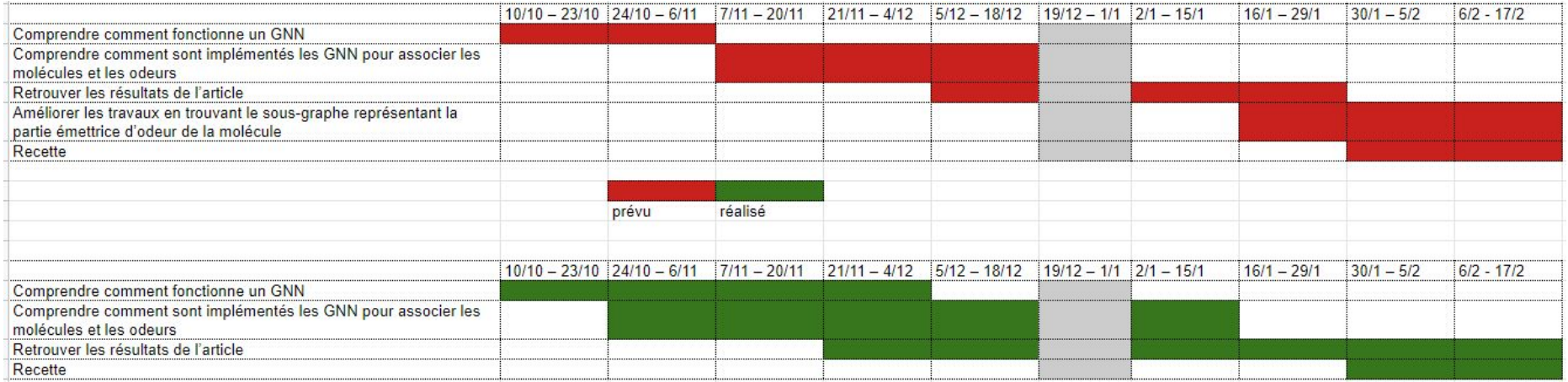


Figure 3 : Gantt prévisionnel et effectif

02. Graph Neural Networks (GNNs)

Qu'est-ce que c'est ?

GNN → obtenir un vecteur représentant la molécule

Morgan fingerprints →

Encodage de la molécule | Prédiction des odeurs

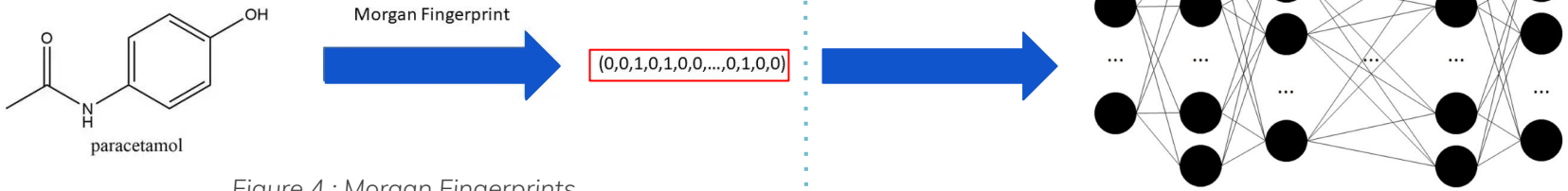


Figure 4 : Morgan Fingerprints

Organisation

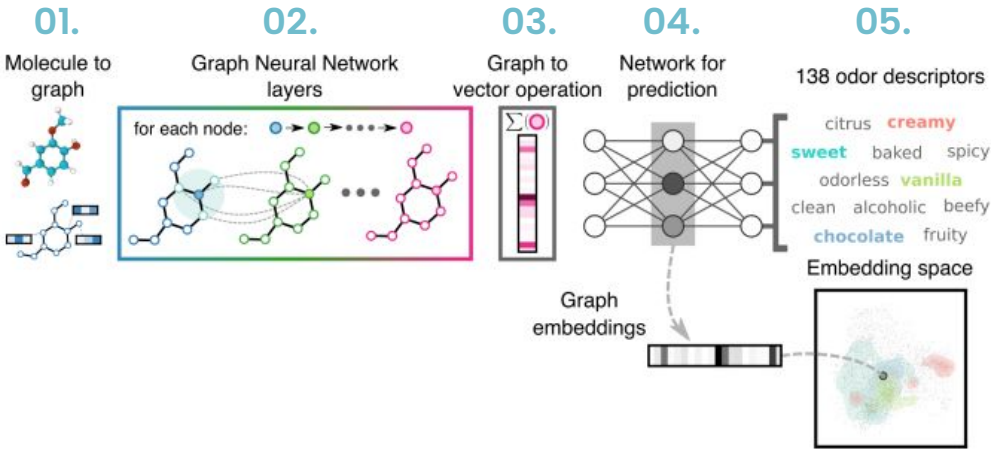


Figure 5 : Modèle schématique de l'article[1] : organisation du GNN

01. Transformation de la molécule en graphe

02. Couches d'échange de messages

03. Vecteur représentant la molécule

04. Réseau de neurones (MLP)

05. Prédiction des odeurs

[1] https://arxiv.org/pdf/1910.10685.pdf

Pourquoi l'utiliser ?

Comparaison du **GNN** avec les méthodes d'encodage :
- bit-based fingerprints (**bFP**)
- count-based fingerprints (**cFP**)

Comparaison du **GNN** avec les modèles :
- random forest (**RF**)
- k-nearest neighbor (**KNN**)

AUROC : l'aire sous la courbe ROC (mesure de la performance d'un classificateur binaire)

Précision : nombre d'observations positives et négatives correctement classées

F1 : combine précision et rappel (recall) en une seule métrique.

	AUROC	Precision	F1
GNN	**0.894 [0.888, 0.902]**	**0.379 [0.351, 0.398]**	**0.360 [0.337, 0.372]**
RF-bFP	0.832 [0.821, 0.842]	0.321 [0.293, 0.339]	0.295 [0.272, 0.308]
RF-cFP	0.845 [0.835, 0.854]	0.315 [0.280, 0.332]	0.295 [0.272, 0.311]
KNN-bFP	0.791 [0.778, 0.803]	0.328 [0.305, 0.347]	0.323 [0.299, 0.335]
KNN-cFP	0.796 [0.785, 0.809]	0.333 [0.307, 0.351]	0.316 [0.292, 0.327]

Figure 6 : Résultats de l'article[1] : comparaison de différents modèles et façons d'encoder la molécule

[1] https://arxiv.org/pdf/1910.10685.pdf

Différentes architectures

Des performances similaires, des architectures différentes

	GCN	MPNN
Message Passing Layers	concatenation message type, 4 layers of dim: [15,20,27,36], selu activation, max graph pooling	edge-conditioned matrix multiply message type, 5 layers of dim 43, GRU-update at each layer
Readout	Global sum pooling with softmax, 175 dim, one per MP layer and summed	Global sum pooling with softmax, 197 dim, one per MP layer with residual connections and summed
fully-connected neural net	2-layers of dim [96, 63] with relu, batchnorm, dropout of 0.47	3-layers of dim 392 with relu, batchnorm, dropout of 0.12 and l1/l2 regularization
Prediction	Multi-headed sigmoid, 138 tasks	
Training	Weighted-cross entropy loss, optimized with Adam, used learning rate decay with warm restarts, 300 epochs	

	AUROC	Precision	Recall	F1
MPNN	0.890 [0.882, 0.898]	0.379 [0.352, 0.399]	0.387 [0.366, 0.408]	0.362 [0.335, 0.375]
GCN	0.894 [0.888, 0.902]	0.379 [0.351, 0.398]	0.390 [0.365, 0.412]	0.360 [0.337, 0.372]

Figure 7 : Performances GCN / MPNN de l'article[1]

[1] https://arxiv.org/pdf/1910.10685.pdf

03. Fonctionnement des GNNs

De la molécule à un graphe

On encode les caractéristiques de chaque atome de la molécule :

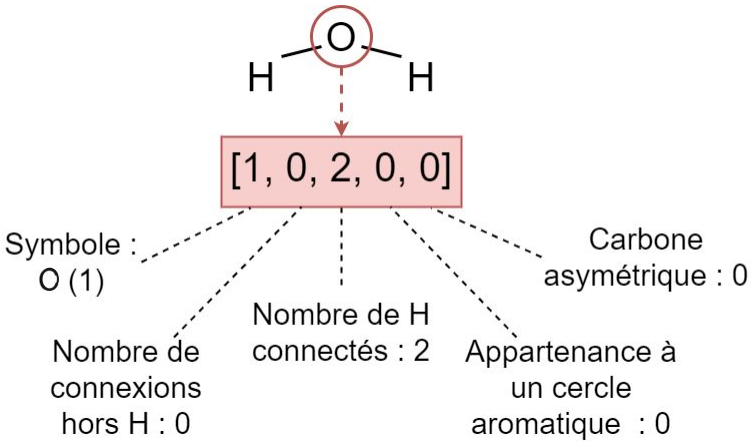


Figure 8 : Encodage de l'atome O de la molécule H_2O

Symbole de l'atome (0 1 2 3 4 5 6) :
(C O N S Cl Br H)

Degré de l'atome (0 1 2 3 4) :
nombre de voisins de l'atome hors hydrogène

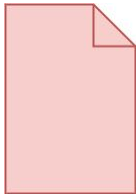
Valence implicite (0 1 2 3 4) :
nombre d'atomes d'hydrogène connectés à l'atome

Noyau aromatique (0 1) :
Appartenance ou non à un noyau aromatique

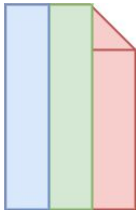
Chiralité (0 1 2) :
(0: non asymétrique, 1: asymétrique sens horaire, 2: asymétrique sans anti-horaire)

L'échange de messages

Pour chaque couche, tous les sommets du graphe (atomes), vont échanger leurs caractéristiques avec leurs voisins



Caractéristiques de **l'atome rouge**



Caractéristiques agrégées des atomes **rouge**, **vert** et **bleu**

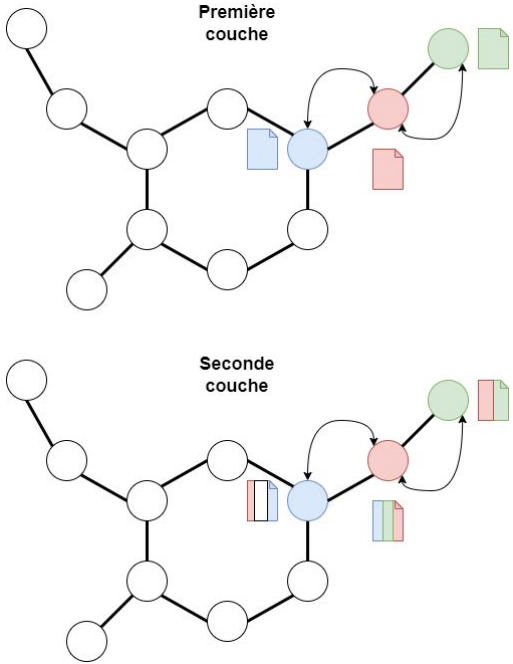


Figure 9 : Exemple d'échange de messages entre l'atome rouge et ses voisins

L'échange de messages

$\bar{A}$ matrice d'adjacence + identité

D matrice des degrés

$\hat{A}$ matrice de régularisation

H^t matrice des caractéristiques à la couche t

W^t matrice des poids à la couche t

H^{t+1} matrice des caractéristiques à la couche t+1

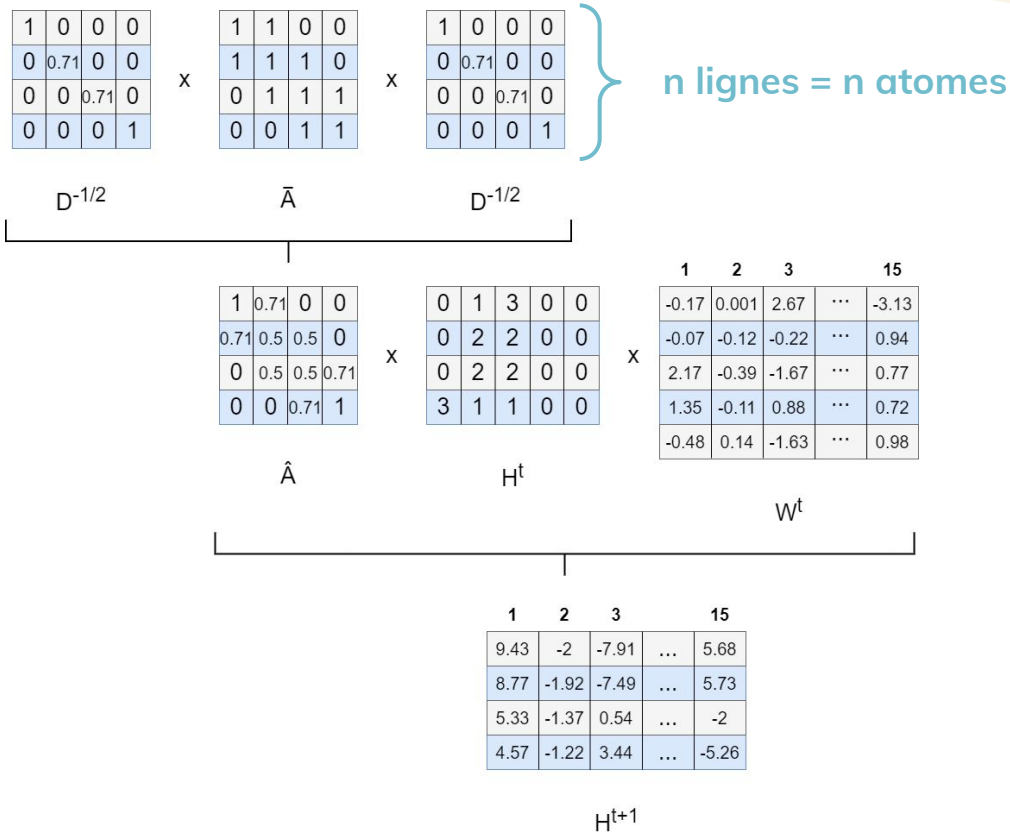


Figure 10 : Détail des calculs du processus d'échange de messages

Sortie du GNN

Matrice de dimension (n, 36) —global pooling→ vecteur de taille (175)

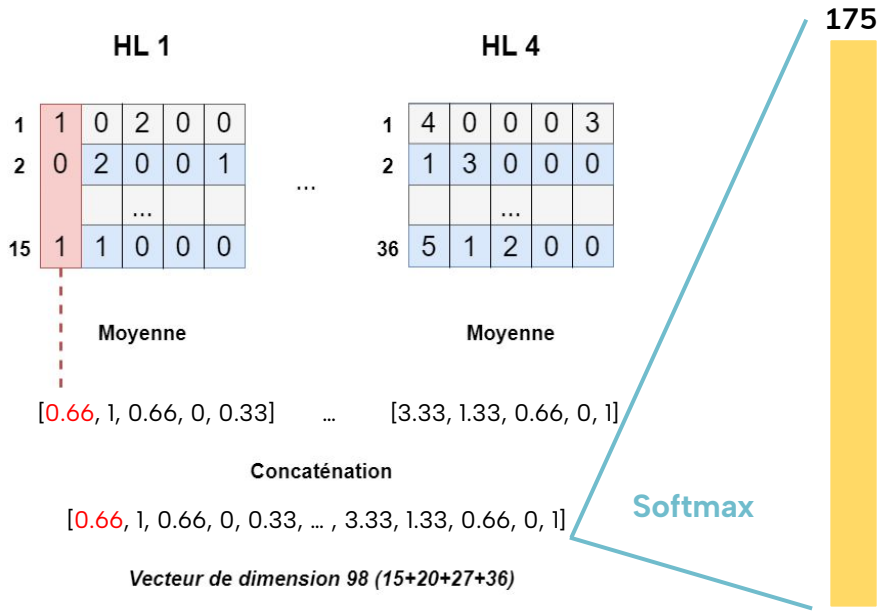
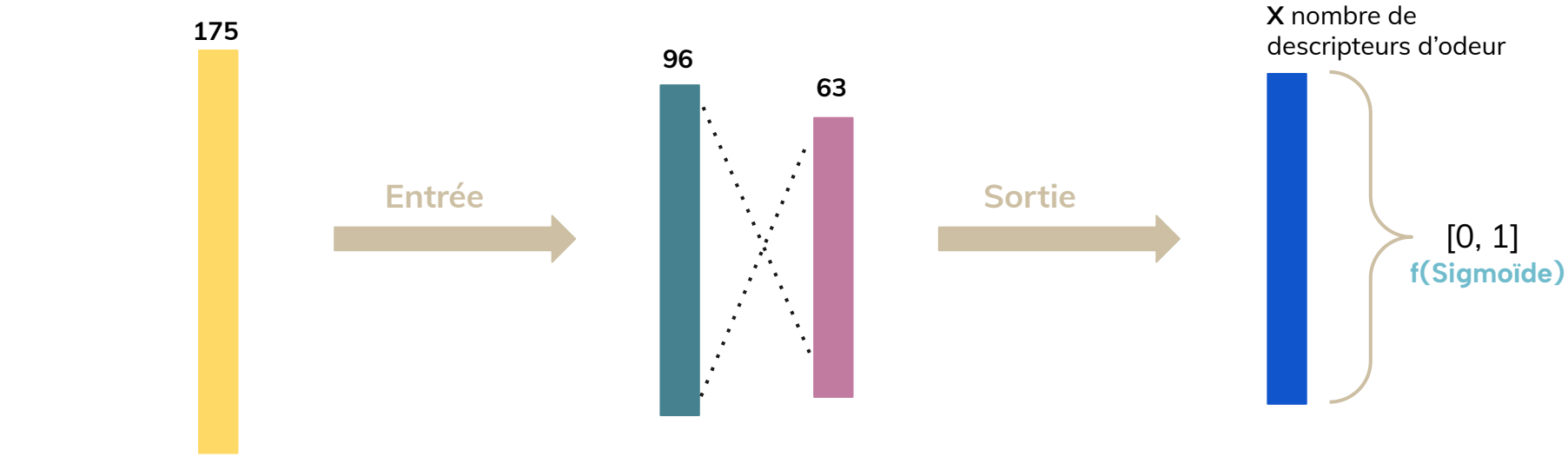


Figure 11 : Détail du fonctionnement du global average pooling

Prédiction des odeurs



Vecteur représentant la molécule

MLP entièrement connecté

Prédiction des odeurs

Architecture entière

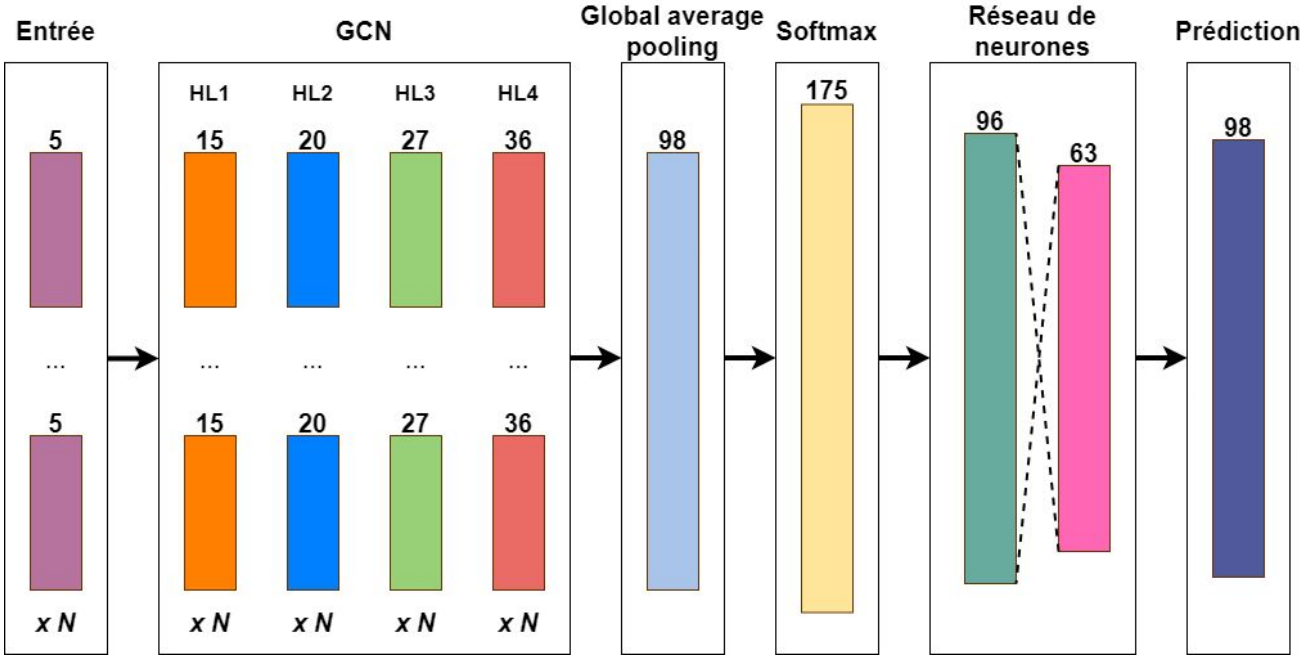
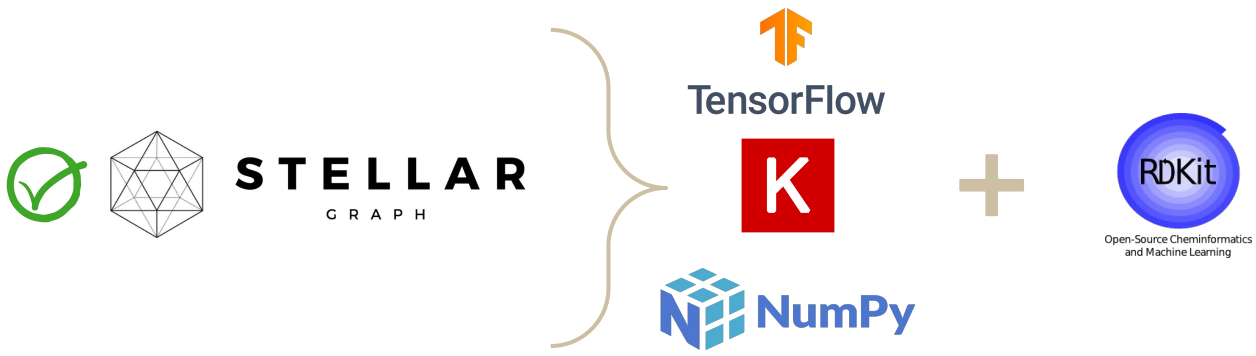


Figure 12 : Architecture du GCN

04. Notre implémentation

Librairies



Les fonctions sont moins transparentes qu'avec StellarGraph

https://stellargraph.readthedocs.io/en/stable/index.html

Données

Données du laboratoire de biochimie d'Oniris

Ensemble de **4 573 molécules** et **381 descripteurs d'odeur**

Données initiales

Prendre les descriptifs d'odeur qui apparaissent **plus de 30 fois**

Ensemble de **98 descripteurs d'odeur**

Tri

Prendre les molécules qui ont un **SMILE** et au moins **1** descripteur d'odeur

Données finales

Ensemble de **2 843 molécules** et **98 descripteurs d'odeur**

Caractérisation des molécules

SMILE → RDKit (Open-Source Cheminformatics and Machine Learning) → **Objet "Molécule"**

Pour chaque atome

Caractéristiques

- **Symbole atomique**
- **Degré**
- **Valence implicite**
- **Aromatique**
- **Chiralité**

\+ **Matrice d'adjacence de la molécule**

`GetSymbol`(*(Atom)arg1*) → str :
Returns the atomic symbol (a string)

`GetDegree`(*(Atom)arg1*) → int :
Returns the degree of the atom in the molecule.

`GetImplicitValence`(*(Atom)arg1*) → int :
Returns the number of implicit Hs on the atom.

`GetChiralTag`(*(Atom)arg1*) → ChiralType :
C++ signature :
RDKit::Atom::ChiralType GetChiralTag(RDKit::Atom {lvalue})

`GetIsAromatic`(*(Atom)arg1*) → bool :
C++ signature :
bool GetIsAromatic(RDKit::Atom {lvalue})

https://www.rdkit.org/docs/GettingStartedInPython.html

Modèle du GCN

```
gc_model = GCNSupervisedGraphClassification(
    layer_sizes=[15, 20, 27, 36],
    activations=["selu", "selu", "selu", "selu"],
    generator=generator,
    pool_all_layers=True
)

x_inp, x_out = gc_model.in_out_tensors()
predictions = Dense(units=96, activation="relu")(x_out)
predictions = BatchNormalization()(predictions)
predictions = Dropout(0.47)(predictions)
predictions = Dense(units=63, activation="relu")(predictions)
predictions = BatchNormalization()(predictions)
predictions = Dropout(0.47)(predictions)
predictions = Dense(units=n_odors, activation="sigmoid")(predictions)
```

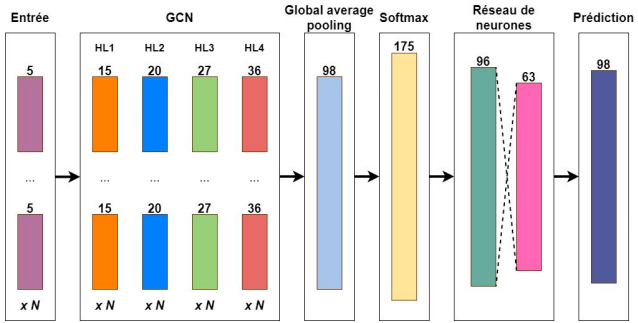


Figure 13 : Architecture du GCN sous Stellargraph

05. Nos résultats

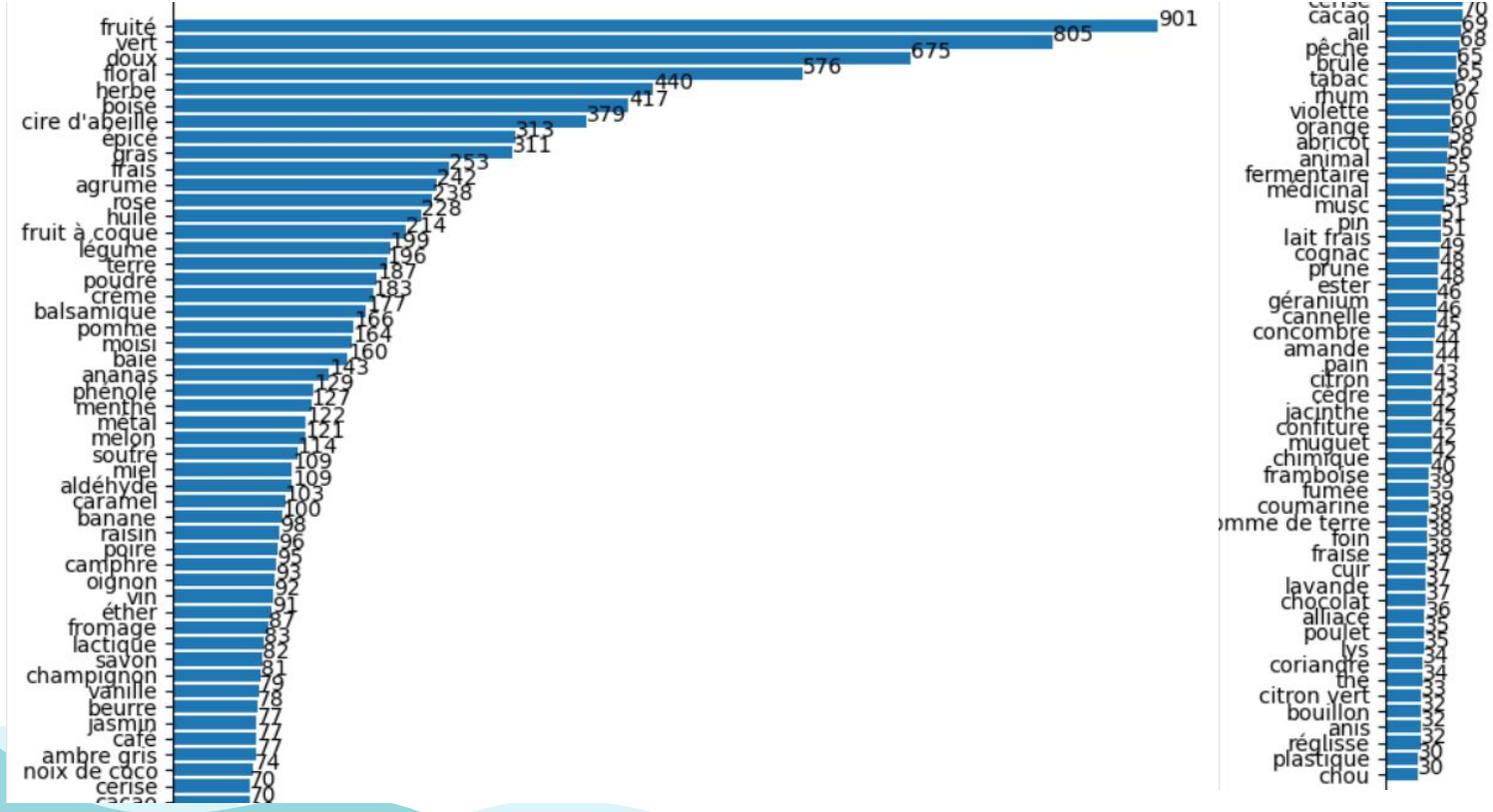


Figure 14 : Fréquence d'apparition des odeurs

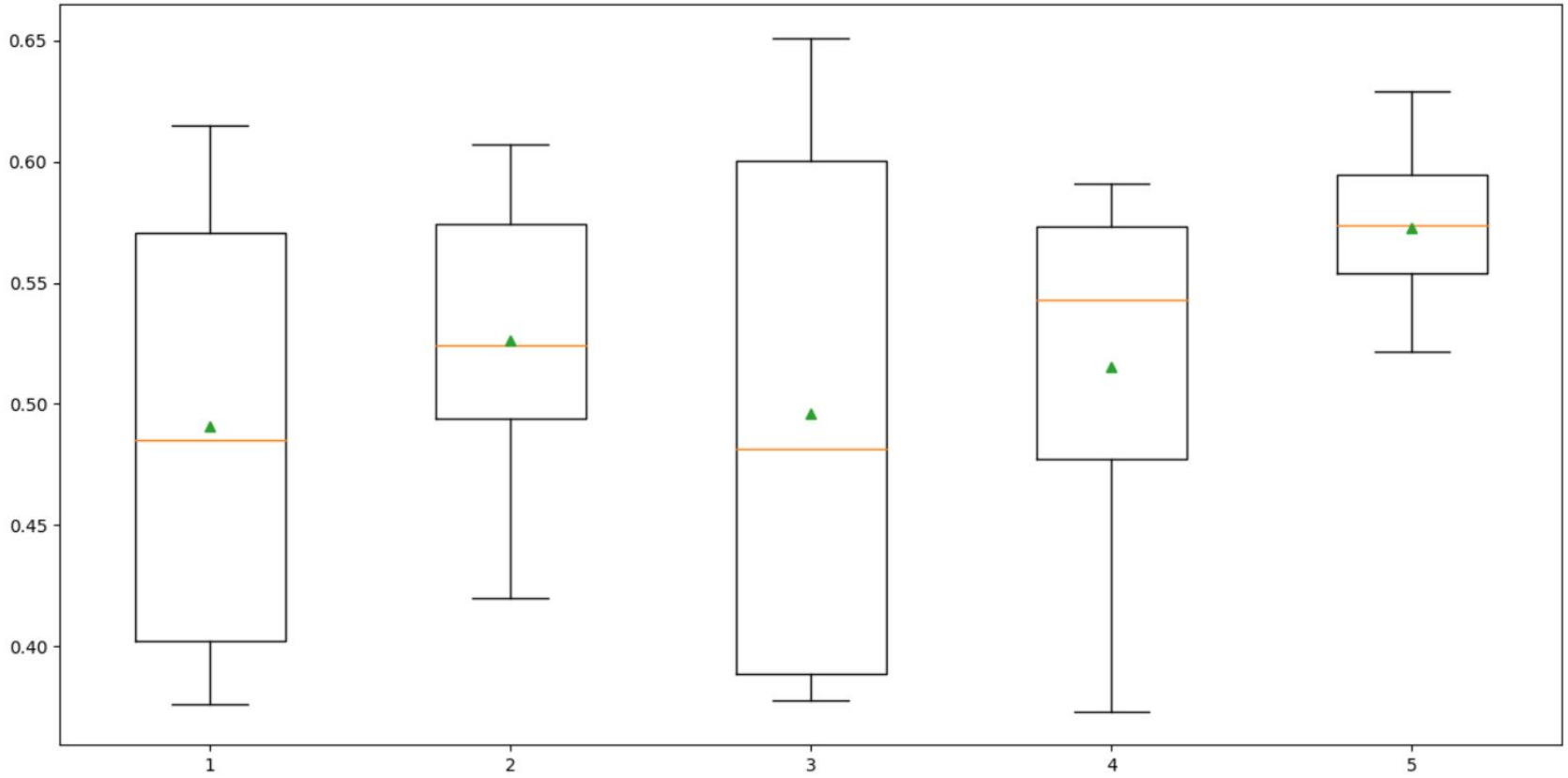


Figure 15 : Précision des différents modèles en fonction du pli

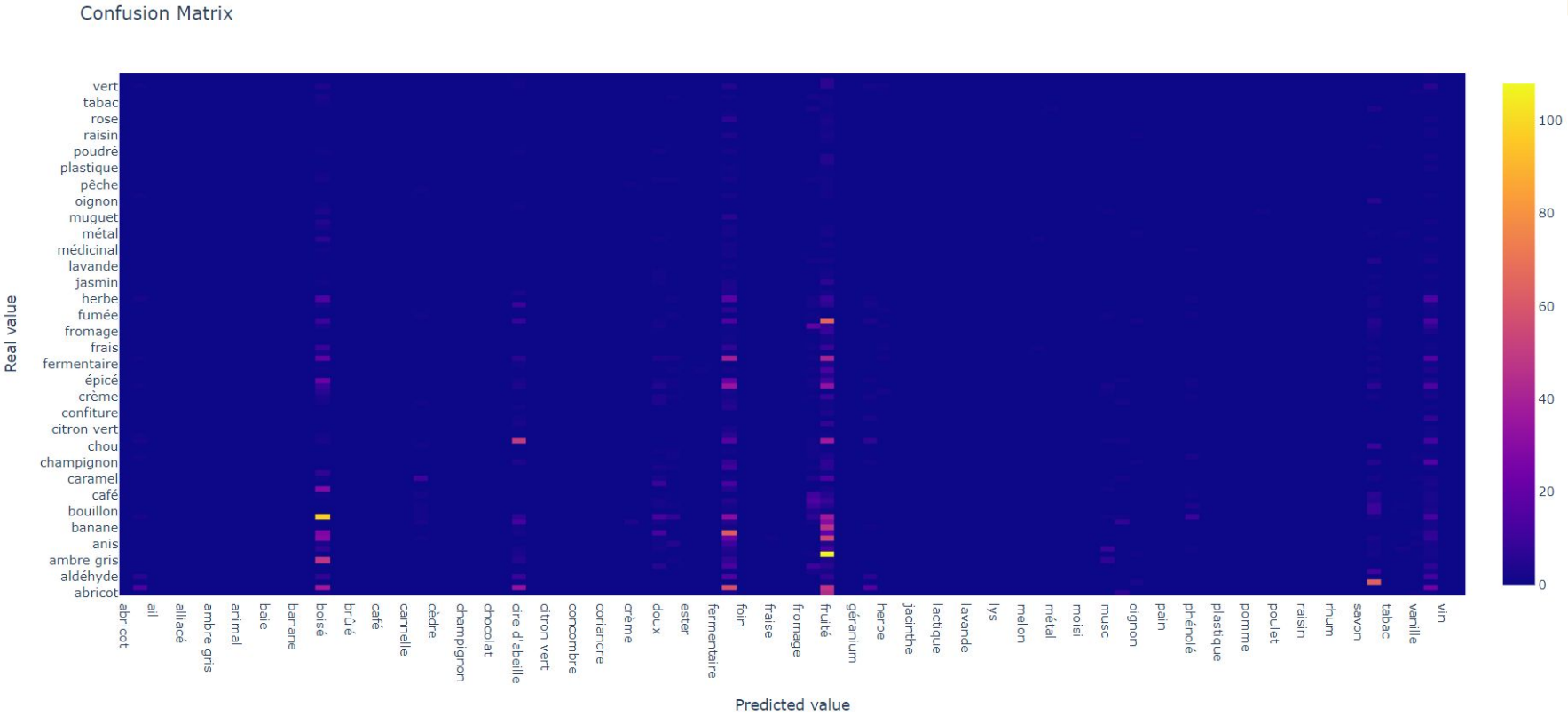


Figure 16 : Matrice de confusion : odeurs prédites / réelles

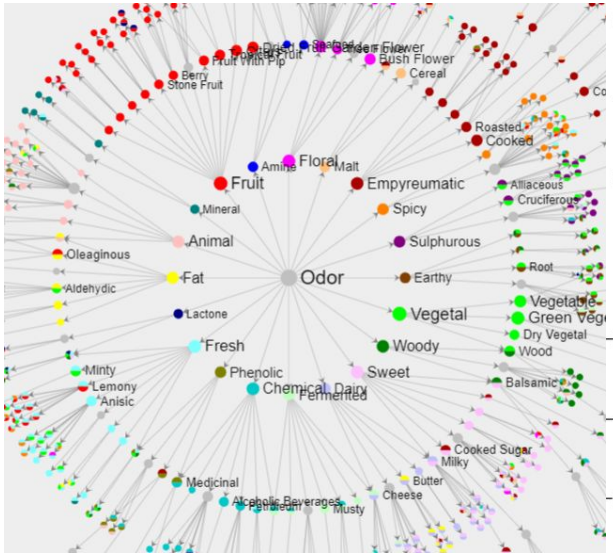


Figure 17 : Ontologie des descripteurs d'odeur[1]

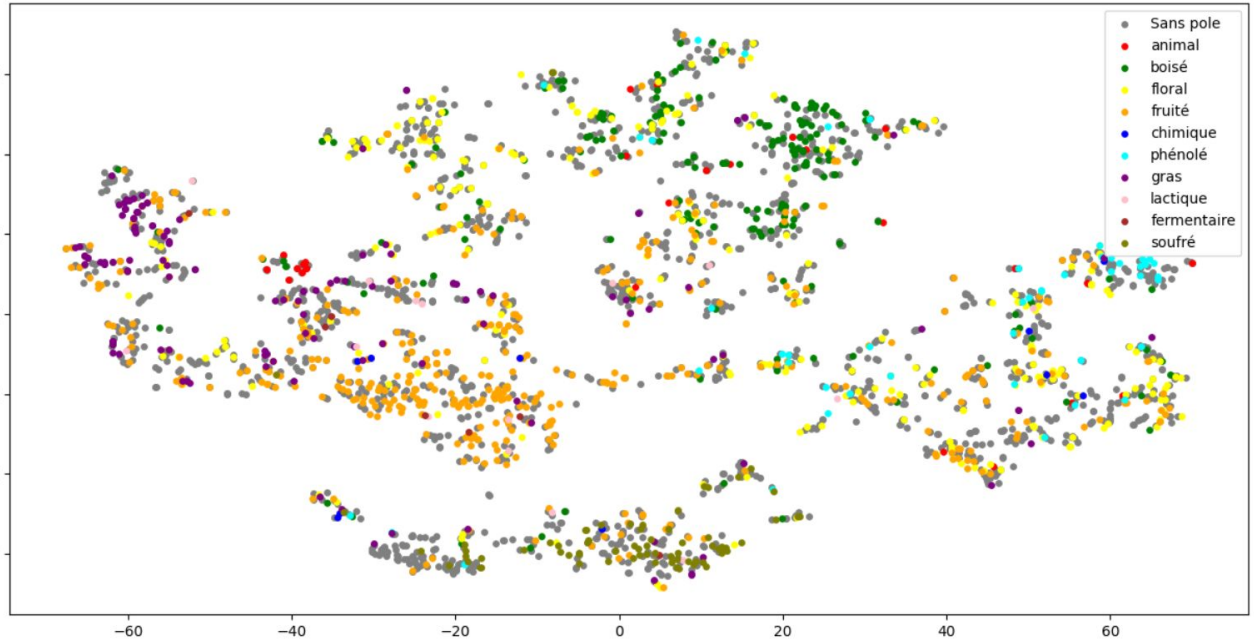


Figure 18 : Représentation 2D des caractéristiques des molécules

[1] https://oniris-polytech.univ-nantes.io/sketchoscent/

06. Conclusion

Nos difficultés

- Compréhension des articles
- Fonctionnement des GNNs
- Comment caractériser une molécule
- Deepchem

Pistes d'améliorations

- Encodage : Type des liaisons
- Retrouver le sous graphe responsable de l'odeur
- t-SNE avec odeurs prédites
- Matrice de confusion en multilabel : définir un seuil (très compliqué)

Bibliographie

Machine Learning for Scent: Learning Generalizable Perceptual Representations of Small Molecules
https://arxiv.org/pdf/1910.10685.pdf

Learning to Smell: Using Deep Learning to Predict the Olfactory Properties of Molecules
https://ai.googleblog.com/2019/10/learning-to-smell-using-deep-learning.html

Odor-GCN: Graph Convolutional Network for Predicting Odor Impressions Based on Molecular Structures
https://assets.researchsquare.com/files/rs-1377643/v1_covered.pdf?c=1667972722

SEMI-SUPERVISED CLASSIFICATION WITH GRAPH CONVOLUTIONAL NETWORKS
https://openreview.net/pdf?id=SJU4ayYgl

Understanding Graph Convolutional Networks for Node Classification
https://towardsdatascience.com/understanding-graph-convolutional-networks-for-node-classification-a2bf db7aba7b

StellarGraph
https://stellargraph.readthedocs.io/en/stable/index.html

RDKit
https://www.rdkit.org/docs/GettingStartedInPython.html

Jure Leskovec
https://cs.stanford.edu/people/jure/teaching.html

Table des figures

[1] https://arxiv.org/pdf/1910.10685.pdf

[2] https://oniris-polytech.univ-nantes.io/sketchoscent/